DOCUMENT RESUME

ED 238 933 TM 840 035

AUTHOR McArthur, David L.; Choppin, Bruce H.
TITLE Evaluating Diagnostic Hypotheses.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 83
GRANT NIE-G-83-0001
NOTE 86p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Adaptive Testing; Computer Assisted Testing; *Diagnostic Tests; *Educational Diagnosis; Elementary Secondary Education; *Evaluation Methods; *Medicine; Models; Performance Factors; Testing

ABSTRACT

This paper examines the history of approaches to diagnosis in education and in medicine--a profession with concentrated attention to the conceptual and mathematical underpinnings of diagnosis. Presented is a comprehensive model of diagnostic testing in education and a summary of the results of four studies, one from each of four separate heuristics developed within the model. The paper concludes with a discussion of the advantages, disadvantages, and possible productive directions for educational diagnosis, particularly in the realm of individualized adaptive diagnostic testing administered by computer. A report "Some Strategies for Constructing and Validating Diagnostic Hypotheses" is appended. (PN)

METHODOLOGY PROJECT DELIVERABLE

EVALUATING DIAGNOSTIC HYPOTHESES

by

David L. McArthur and Bruce H. Choppin
Project Directors

Grant Number
NIE-G-83-0001

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California
November, 1983

SOME STRATEGIES FOR CONSTRUCTING AND VALIDATING DIAGNOSTIC HYPOTHESES

by

Bruce H. Choppin
and
David L. McArthur
Project Directors

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Table of Contents

Table of Figures

EVALUATING DIAGNOSTIC HYPOTHESES

by

David L. McArthur
and
Bruce H. Choppin

Center for the Study of Evaluation

UCLA

I. INTRODUCTION

Diagnostic testing in education is undergoing a revolution. On one hand a fair number of specialized test protocols are extant which are called "diagnostic," a large quantity of statistical and psychometric theory can be applied to the diagnostic question, and computer technology promises to deliver into the hands of the classroom teacher systems which will teach, test, diagnose and remediate a variety of educational offerings. On the other hand, diagnostic testing in most areas of education builds on weak theoretical foundations, makes use of few statistical tools and none of the wealth of experience available from diagnostic testing in other professions, and with rare exceptions does not yet draw on the power of computers.

This paper examines the history of approaches to diagnosis in education, and in a profession with far more concentrated attention to the conceptual and mathematical underpinnings of diagnosis, the field of medicine. We present a comprehensive model of diagnostic testing in education and a summary of the results of four studies, one from each of four separate heuristics developed within the model. The paper concludes with a discussion of the advantages, disadvantages, and possible productive directions for educational diagnosis, particularly in the realm of individualized adaptive diagnostic testing administered by computer.

The phrase "diagnostic testing" has been used in education ever since the first formal intelligence tests were devised. From the beginning the nominal intent of educational diagnosticians appears to have been relatively stable: "...the taking of certain symptoms that exist and finding out from them what the trouble is" (Kallom, 1919, p. 11). While the diagnoses themselves, the process by which diagnosis is reached, and the management decisions which follow have undergone numerous and extensive revisions; whether to build on a disease model or some alternative such as learning theory has been a constant source of controversy. Except in reference to specialized psycho-educational and physical handicaps, however, with few exceptions there is not a great deal to show for the effort (Tyler & White, 1979)

The common thread behind most approaches to educational diagnosis in the past seven decades has been the use of tests to provide

specific information about the difficulties of an individual student which will point to some appropriate remedial treatment. The phrase "diagnostic testing" increasingly is being used for the assessment of learning difficulties within the classroom. In order to arrive at a diagnosis of individual patterns, existing tests which use the diagnostic label diverge widely in their approach, yet all are concerned to some degree with the following key elements:

a) examination of patterns of performance and achievement of an individual student,

b) construction of a summary profile of strengths and weaknesses,

c) identification of the specific misunderstanding, misconceptions, and misinformation that lead the individual student to perform poorly.

Viewed in this manner, diagnosis of difficulties experienced by an individual student could lead to appropriate management strategies for further learning, remediation, re-education, or referral.

The earliest efforts at developing diagnostic strategies in education were predicated on a very similar rationale. Uhl's (1917) diagnostic method emphasized close examination of each pupil's methods of work and questioning of students aloud while they solved a problem. Uhl developed a series of hypotheses concerning students' incorrect methods and recommended drilling pupils in methods which are "more effective" than those they already employ. Anderson (1918) discussed diagnostic testing in reference to seven types of errors in long division. Subjects were given individual oral tests in which

they were asked to think aloud and to say what they were thinking and doing while solving the problem. Anderson's aim was to enable teachers to become diagnosticians of "mathematical diseases." Paulu's (1924) Diagnostic Testing and Remedial Teaching gave numerous examples of tests in spelling, writing, reading, arithmetic, geography and history which had diagnostic potential. Paulu urged that teachers observed their students' working procedures and learn to recognize individual difficulties. The number of times each problem was incorrectly solved, the body movements made by the child while working, and the use of finger-counting were a few of the examples viewed as important signs of difficulty to be followed by specific individual remediation.

The first volume of Journal of Educational Research contained a study of diagnosis of error types (Willing, 1920); the first volume of Journal of General Psychology contained a major article by Spearman (1928) on the "Origin of error"; the second volume of the British Journal of Educational Psychology presented a lengthy analysis of theories of cognitive error (Fortes, 1932). In general, errors tend to show themselves as matters of either principle (such as faulty reasoning, misunderstanding, or inability to apply a correct method or strategy) or accuracy (such as errors in copying, manipulating numbers, or misplacing parts of the problem).

While these historical documents present a minimum of sophisticated conceptualization, the present status of many

application-oriented publications in educational diagnosis is not many steps further. Despite the intention to make use of charts, graphs, and profile analysis, how-to books like Smith (1969), articles like Okey (1976), and computer programs like Furlong and Miller (1978), for example, contain relatively few substantive advances in either the specificity of diagnosis or the range of options available to teachers in both developing and utilizing a given diagnostic test. Moreover, two essential definitions often appear absent from diagnostic tests and manuals. The first is the meaning of the word "pattern;" a number of sources use this word but its meaning varies rather widely:

a) "pattern" as profile of total scores in a curricular domain accumulated across a variety of tests administered throughout the school year ("a pattern of deficient test scores in spelling");

b) "pattern" as profile of subscale scores assembled from a single test administered once ("a pattern of misunderstanding of two-digit arithmetic");

c) "pattern" as consistent behaviors across differing situations ("a pattern of hyperactivity");

d) "pattern" as unusual responses to a set of test items ("a pattern of responses which points to carelessness on this test");

e) "pattern" as specific erroneous responses within a set of test items ("a pattern of responses which demonstrate consistent errors in logic").

The various writers do not appear to have thought that "pattern" raises such plethora of possibilities. The list is not exhaustive, nor are the entries mutually exclusive, but often defficient. Recommendations for interpreting "patterns," hinge on the reader's correct choice of definition.

The second word requiring definition, surprisingly, is "diagnosis" itself. A number of educational writers refer to "diagnosis" as if it is either perhaps self-explanatory or too trivial to discuss, yet an adequate definition of the term is critical for purposes of further refinement and application.

As effort is expended in developing and administering diagnostic tests of increasing sophistication, an improved definition of diagnosis can be developed by examining the range of present applications, the state of theory concerning diagnosis, and contributions from the field of medicine. The following sections address each in turn.

II. VARIETIES OF DIAGNOSTIC TESTING IN EDUCATION

1. Testing ability vs. testing achievement

The vast majority of tests in education today may be grouped into one of two categories: (a) specific or general ability tests (e.g., intelligence tests) designed to measure a student's innate ability or potential, and (b) achievement tests designed to measure how much a student has learned.

In a sense, almost any standardized test of ability or achievement may be regarded as diagnostic. But the practice of educational testing has broken into distinct categories, of which the major ones are placement and selection, grading and certification, motivation and research as well as diagnosis. The placement and selection operations grew directly from the work of Alfred Binet and the extensive use of objective intelligence tests during World War I

in the evaluation and placement of new recruits. Soon a wide variety of intelligence and achievement tests were being made available to employers and to technical and vocational training institutions for the purposes of screening new applicants, and both achievement and intelligence tests are in worldwide use today for placement. Following the meaning of "pattern" as a profile of subscale scores, placement and selection tests are "diagnostic" in the sense that a pattern of test profiles may be used for differential assignments.

Objective tests are widely used to study various aggregate aspects of the educational process. This category of use encompasses measurements embedded in the design of educational experiments, the evaluation of new educational programs or curricula, and the monitoring of district, state, or national levels of achievement. Achievement tests are used extensively to measure outcomes and hence to "diagnose" the effectiveness of instructional programs, specific school districts, or individual teachers. The use of objective tests in the certification process at the end of a specific program of education or training is seen primarily as a method of maintaining standards over time and are often only crude diagnostic indicators. Since changes in general ability or intelligence are thought to be mostly beyond the scope of the educational system, the use of ability tests in a research or evaluation setting is usually not treated diagnostically, but rather as a way of controlling the experimental design or of "explaining" away some of the observed variance of achievement scores.

Diagnostic testing of the individual student is the category with which this paper is primarily concerned. In the same way that the psychologist has a wide range of diagnostic ability measures to aid the identification of various sensory defects or brain dysfunctions, now the teacher has access to "diagnostic" tests as well. While the notion of diagnostic achievement testing has been around for several decades, the appearance of large numbers of objective achievement tests which purport to be diagnostic is a recent phenonemon. Clearly a model of the diagnostic process which translates directly to the classroom setting and needs of the teacher to diagnose education problems would aid in understanding and utilizing the range of tests. One model which begins to meet these needs is provided by Thomas (1983) and is presented below.

It is useful to distinguish between testing for specific learning disabilities and more general assessments of learning achievement. Hennessy (1981) points out that the primary use of individually administered tests in schools today "is to obtain descriptions of functioning for the purpose of diagnosis of children thought to be learning disabled, neurologically impaired, developmentally disabled, or emotionally disturbed" (p. 42). Indeed codes of practice in many states require that individually administered abilities measures shall be included as part of the diagnosis of children prior to their classification or assignment to special educational programs.

A variety of different conditions are subsumed under the general title specific learning disabilities, and there seems little doubt that many of these conditions do result from, or are related to, particular brain malfunction or damage. A variety of psychological tests have been devised to assist their identification, though Arter and Jenkins (1979) and Hennessy (1981) point out how limited the evidence of validity is for these tests. However, such tests are generally the prerogative of the trained clinical psychologist, and are not customarily used by (and may not be legally available to) the classroom teacher. But the classroom teacher's needs are not identical: frequently the task is not one of locating disability or disturbance but rather one of finding and understanding where a student has encountered a block, is using an erroneous strategy, or has been otherwise left by the wayside.

2. Testing and diagnosing individual educational performance

Thomas (1983) distinguishes between diagnostic and other forms of evaluation in terms of the sort of question each addresses and the uses typically made of the evaluation data:

> "With diagnostic evaluation, the question consists of two parts: what is the pattern of strengths and weaknesses in the students' achievement of the learning goals, and what causes underly such a pattern? Results of such diagnosis are used for recommending treatment of a student's learning weaknesses, either through remediation of underlying causes or through helping the pupil learn more adequately despite the causes." (p. 13)

Basic to this approach to diagnosis is the interpretation of the pattern of performance scores. Here the use of the word "pattern" can

be interpreted both as a profile of subscale scores and as consistent or unusual references and behaviors. Although it is not essential, often such patterns are derived by comparing an individual's performance with that to be expected based on the results of some reference group, an approach may appropriately be described as "norm-referenced".

Thomas recommends a methodical approach to the diagnostic use of tests, whether by classroom teacher or school psychologist. He points out the errors that can result from steps being omitted and short cuts being taken. For example, a very poor reading performance as measured on a general abilities test may stem from any one of a variety of completely unrelated causes, and further investigation is necessary before appropriate treatment can be confidently prescribed.

Thomas' approach to diagnostic assessment of students, shown in Figure 1, is comprehensive although time consuming. It succeeds in codifying what teachers are supposed to be doing when they provide individualized instruction. The model is not limited to the norm-referenced approach and may also be applied to criterion-referenced testing, as will be discussed below. Furthermore, it may succeed in identifying and diagnosing the causes of major problems, although it is less likely to be sensitive to specific misunderstandings, misconceptions, and misinformation which may be significant to an individual student in his mastery of a given topic.

Figure 1
The Thomas Diagnostic Evaluation Model

Stage 1 : Status Assessment

Critical questions :
- 1.1 What are the specific objectives the student is expected to have achieved?
- 1.2 What assessment techniques can best determine how well the student has achieved those objectives?
- 1.3 What pattern of discrepancies between expectations and performance is identified by these techniques?

Stage 2 : Cause Estimation

Critical questions:
- 2.1 What reasons for the deficiencies revealed in 1.3 need to be considered?
- 2.2 How can these possibilities be evaluated?
- 2.3 On the basis of these evaluations, what is the most likely cause (or combination of causes) for the pattern in 1.3?

Stage 3 : Treatment

Critical questions:
- 3.1 What treatments would help the student most effectively given 1.3 and 2.3?
- 3.2 What evaluation techniques are available to determine how well the treatment is succeeding?
- 3.3 As assessed by these techniques, how successful is the treatment?

(After Thomas [1981], p. 15-16)

An increasing number of commercially published standardized achievement tests are now incorporating the "diagnostic" label into their title. However, it would seem hard to justify the label for any test that produces only a single score. Not only do such tests provide no indications of the likely cause of a particular result and no suggestions as to appropriate remedial treatment (as required by Thomas' model), but the single score can be only a small part of the data needed to build up the pattern on which normative diagnosis rests. A reading comprehension test may indicate, with high reliablility and validity, that a sixth grade student is reading at the fourth grade level, but the information needed for diagnosis of the student's problems would not be found unless some detail such as subscale scores or specific erroneous response patterns is also made available. It would be more reasonable to reserve the term diagnostic for batteries of standardized tests which yield fairly complete profiles of performance in normative terms--the interpretation of which might well suggest both causal factors and remedial treatments.

Such patterns of scores, or normative profiles, are very important in norm-referenced diagnostic testing. A key issue for the practitioner is the level of detail on which the components of the profile are differentiated. Component elements of three different profiles produced by three hypothetical diagnostic test batteries might be:

General Achievement Tests

Reading comprehension, Handwriting, Math skills, Social Studies concepts, Science facts and concepts.

Mathematics Test

Computation skills, Fractions, Numerical reasoning, Algebraic manipulation, Geometric similarity and congruence

Magnetism Test

Magnetic and non-magnetic materials, Magnetic attraction and repulsion, Concept of a magnetic pole, Induced magnetism, Concept of a magnetic field, The Earth's magnetic field.

Although each relies on the same underlying theory, the interpretation of results and the prescription of remedial treatment would be quite different in each case. The first example gives only global information but might be helpful in indicating whether or not a student's problems stem from a perception problem, a linguistic difficulty, or some type of specific learning disability with physiological roots. By contrast, the second list of profile components will be chiefly useful in indicating areas of instruction which have not been mastered by the student, due to some dislocation of the normal teaching/learning process. For students with very discrepant patterns it may indicate a need for substantial remedial study.

The third list of profile components represents an assessment of performance objective-by-objective. While this might appear the most useful form of assessment for detailed implementation of an instructional program, it must be recognized that a great deal of time is required to obtain reliable estimates of individual profiles at this level of detail. By aggregating the results of just a few items

across the students in a class, a teacher quite economically can obtain feedback as to how well the class has mastered specific objectives, information helpful in planning the next step of the teaching sequence. However, this approach does not often provide useful information at the individual level.

In each case scores on the component parts of diagnostic profile may be interpreted as deviations from the norm. Notice, however, that "norms" are established by averaging the scores for large numbers of students, and this does not imply that a flat profile, indicating even levels of development, is to be expected for any or all students. The achievement of most children does not proceed in an orderly and regular fashion, and we should not expect to find unchanging scores as we move from one area to another. Nevertheless, experience suggests that substantial unevenness of development [say two grade levels between subject areas] likely indicates more than a passing disaffection with one subject or another, and further investigation would be appropriate. Components of diagnostic profiles within a particular curriculum area may be expected to be more closely related, particularly if there are strong logical connections between sub-areas, as in mathematics. Even so, the typical student will do better in some areas than in others, and unless the differences are extreme, a serious learning problem is not necessarily indicated. For diagnosis of learning objective-by-objective, norm-referenced interpretations have limited utility. This type of diagnostic battery

is more effective if it can be interpreted in criterion-referenced terms--especially if the sequence and structure of objectives is supported by cognitive learning theory.

Both Thomas (1983) and Hunter (1979) stress the importance of accumulating a wide variety of evidence upon which to base an educational diagnosis. Test scores by themselves can be misleading unless considered in the context of the conditions under which they were obtained, the past performance of the student under consideration, scores of pupils of similar maturity who have been exposed to similar instruction, information about the student's linguistic background, etc. For example, while it is entirely proper that test scores form a part of the data on which any important classification or assignment of a student to a special educational program is based, test scores should not be used alone for such purposes, but should always be supplemented by appropriate contextual information. Likewise, test scores are one of many sources of information upon which a teacher draws in making instructional decisions.

[On the other hand, the use of test scores by an individual for self-diagnosis may be quite effective. The student can integrate diagnostic feedback if appropriately presented with past experience in order to help determine what topics or principles he needs to study more carefully. More research on this type of self-directed learning is needed.]

The important distinguishing characteristic of norm-referenced testing -- the determination of detailed profiles -- rests heavily not only on the reliability of the particular test and its administration, but also on the demonstrable validity of the reference norms, and the implicit assumption that normed profiles, which are composites of many individual profiles, honestly reflect a developmental reality. Few children proceed with their education in an orderly and regular fashion; we should not expect to find unchanging scores as we move from one area to another. Even within a single domain, the typical student performs better in some areas than others. Thus, the norm-referenced approach to diagnostic testing has shortcomings which are difficult to surmount.

In brief, the norm-referenced approach to diagnostic testing has two major shortcomings. The first is the question of the relevance of any particular set of norms to the student being tested, a question easy to raise but not to resolve in the vast majority of cases. The second problem concerns the large number of test items which must be used if reliable and detailed objective diagnostic profiles are to be developed. Can these problems be avoided by switching to a criterion-referenced approach?

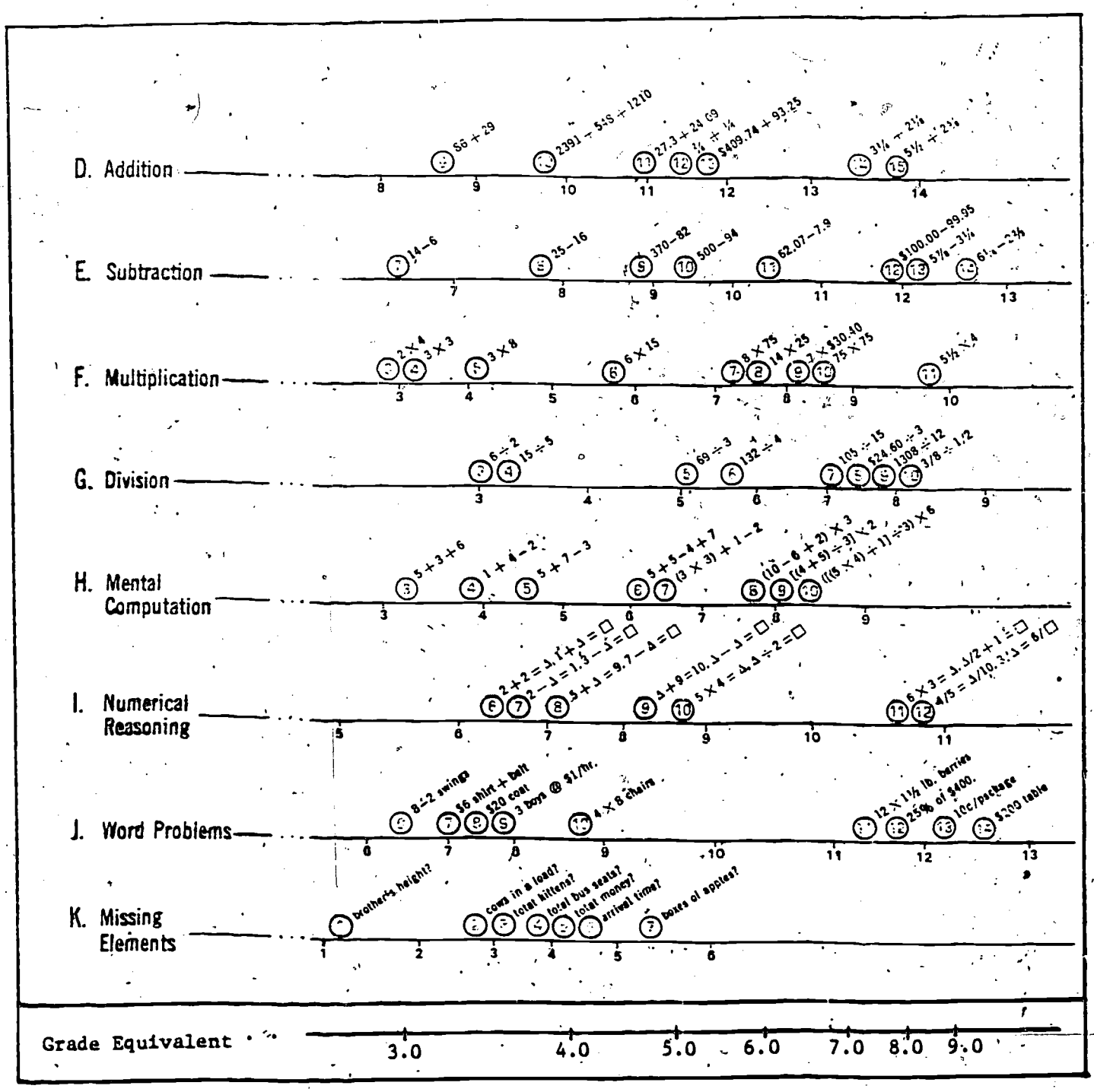
Criterion-referenced tests explicitly attempt to indicate what performances should be expected for students with a given score, without refering to the scores of any other student. In such tests, the issues of relevance to norms and ranking of students are traded

for an issue about the adequacy of items in relation to the criteria being used. A good criterion-referenced test will generate a large amount of information about the overall achievement of a student even from a small number of test items. Because we do not need to relate the pattern of performance for an individual to that of a large normative group in criterion-referenced testing, the testing procedure itself can be made more flexible; an individual student need not attempt all items. Adaptive testing, in which the sequence of items presented to a student depends upon the student's previous responses, offers a much more efficient way of gathering information about the student's achievement and may reduce substantially the time needed to develop a reliable profile (Green, 1983).

A good example of a diagnostic test that uses this adaptive approach is the KeyMath Diagnostic Arithmetic Test (Connolly, Nachtman, & Pritchett, 1971). This test lies somewhere between the pure criterion-referenced and norm-referenced approaches since it has elements of both within its design. The entire instrument consists of 209 test items divided into 14 different components of a diagnostic profile. The diagnostic profile is developed on a large sheet which effectively provides a map of arithmetic attainment with the different content areas listed down the page and the item difficulty levels moving from "easy" on the left to "difficult" on the right. An extract from the complete profile sheet is presented in Figure 2. The circled numbers represent the position of particular items on the

Figure 2.

Scaled summary profile of performance from the KeyMath Diagnostic Arithmetic Test



different scales. Items of equal difficulty appear vertically above or below one another. Scaling according to the Rasch latent-trait model is used to establish these relative difficulties so they form a set of relationships expected to be valid for all students, and not only those belonging to a particular normative group. However, a "grade equivalent" scale is also provided on the diagnostic sheet so that normative interpretations of performance are possible. The strength of this system is that it is adaptive to the needs of the individual student.

3. Analyzing errors

For most of its history, achievement testing has been dominated by the "number correct" method of scoring, so little attention has been paid to the nature of the erroneous responses given by students. Where mistakes have been studied it is to award partial credit for an answer to an open-ended question that was nearly correct (for example, in Great Britain), or for choice of the least incorrect distractor to a multiple choice item (chiefly in the United States). Although both teachers and measurement specialists usually agree that incorrect test responses contain diagnostic information about the student's performance, there have been few systematic attempts to exploit this information.

The advent of computer technology in recent years has led to several attempts at redressing this situation. For example, Brown and Burton (1978) developed the "BUGGY" system, a computerized game for

training teachers in diagnostic skills, which plays the role of a student answering questions. The teacher's task is to recognize the source of the student (computer) error, and to become more sensitive to the causes of students' learning problems. Under this system, simple types of error or "bugs" can be easily diagnosed, although diagnosis becomes much more difficult when the student has several bugs which may interact.

In setting up such a system, the initial identification of misconceptions, or bugs, that produce errors is a complex task. It requires the analysis of each skill under study and of the "procedural network" of subskills, and a listing of the correct and incorrect procedures for applying each of these. In the view of Brown and Burton, this network analysis needs to be comprehensive for it must contain all possible misunderstandings. The need to be comprehensive restricted Brown and Burton to the rather narrow task of addition and subtraction. Even within this field, the number of bugs to be considered is quite large.

This approach has been further elaborated by K.K. Tatsuoka and colleagues at the University of Illinois (Birenbaum & Tatsuoka, 1980; Tatsuoka, et al., 1980; Tatsuoka & Tatsuoka, 1983). They have also concentrated on skills of addition and subtraction of signed numbers using open-ended questions. A major concern of this group was that students might obtain the right answer to a question by applying incorrect reasoning, so that the simple "number right" score on a test

might be an inaccurate indication of achievement. By careful structuring of test questions, they showed that it was possible to infer when a student was using incorrect rules to obtain the correct answer to a specific item by an analysis of responses to other items. Revised scores were produced by rescoring as "incorrect" any correct response deduced to have been reached by wrong reasoning and the revised scores were shown to be superior on each of a number of measurement criteria. This research also demonstrated the inadequacy of factor analysis as a technique for investigating the structure of achievement tests. To a significant extent, the factor structure appears to be determined by the pattern of misconceptions held by the students as well as by the content of the items themselves. Tatsuoka, et al. (1980) introduced the "Individual Consistency Index" (ICI) which, when applied to the pattern of responses for an individual, can indicate the extent to which the student is using "erroneous rules" to solve the problems. However, as pointed out by Tatsuoka and Tatsuoka (1983), since most tests do not have the special structure required for the calculation of the ICI, the method has a limited application.

The detailed analysis that was required to produce a workable system in signed number arithmetic suggests that there will be no general all-purpose computer program that will be able to magically diagnose a pupil's erroneous answers to test items regardless of the subject matter. A full analysis of the logical steps in problem solution outside the area of mathematics is likely to be beyond the

capabilities of teachers, curriculum specialists, and professional test constructors.

However, Nesbit (1966) did demonstrate an approach to diagnostic testing that teachers could handle. It requires that a teacher catalogues the important errors and/or misconceptions common among students in a particular curriculum subdomain, and then writes multiple choice items in which the incorrect alternatives (or distractors) reflect these common misconceptions. Simple analysis of the responses to a set of such questions can indicate whether a student is operating under a particular misconception or not. Though far less comprehensive than the Tatsuoka and Tatsuoka system, and not based on a detailed logical analysis, this approach appears to be much more practical. Even so, experience suggests that the cataloging of error types in a way that multiple choice questions can differentiate between them still requires considerable preparation, and groups of teachers working together may find this more feasible than individual teachers. The use of multiple choice rather than open-ended questions has the disadvantage of denying a student with an unusual misconception or erroneous rule from the opportunity to demonstrate it, but does focus on the main or most frequently encountered errors.

III. DIAGNOSIS IN MEDICINE

While diagnosis in many areas of education has been making rather slow progress, the last dozen years have seen enormous growth in theory and practice in the field of medicine. The limitations which

prevent wholesale borrowing from medical applications are both obvious (does a "disease" model apply to educational difficulties?) and subtle (does educational measurement achieve equal probabilistic accuracies?). However, recent developments within medicine, and especially within the technical field of artifical intelligence in medical diagnosis, make it important for educator's to examine the sucesses and failure even if details of the diagnostic question are not completely parallel between professions.

The practice of diagnostic medicine has been under refinement for as long as medical schools have existed in America. Indeed for an extended period of time, except for a few medicinals and a limited surgical repertoire, the practice of medicine was virtually restricted to the formulation of diagnoses. In this one area, physicians were able to develop extensive and often labyrinthian categories within categories, developing and occasionally discarding the pieces of diagnostic nosology, building a foundation of modern diagnostic practice. Today's general practitioner faces thousands of possible, fully legitimated, diagnostic situations; for the common kinds of illness, all of the following are likely to be true:

a) the category is a recognized and documented disease entity;

b) the status indicators - signs, symptoms, and relevant history, are either specifically understood and delineated or, at worst, have already undergone detailed study;

c) the probabilities associating the presenting symptoms with a variety of disease hypotheses are known fairly closely;

d) the probabilities associating the various diseases with best-fitted therapeutic strategies are known in a general way;

e) the probabilities associating each disease and its recommended treatment with patient outcome are at least roughly estimated;

f) the various combinations of costs and benefits, including situations in which two or more diseases are compounded with each other in the same case, are calculable.

Thus many presenting patient problems can often be translated by cookbook into unambiguous terms: the medical problem is "x" within a specific confidence interval, its course is fully anticipated with (and without) treatment "y" and such treatment has a closely predictable likelihood of benefit at a known cost.

From such a highly defined diagnostic structure has emerged a variety of sophisticated models used to explain the manner in which the professional enters and exits the diagnostic question, how the various paths are profitably explored, and how the disease entity, in time, is understood both statically and dynamically (Gheorghe, Bali, Hill & Carson, 1976; Miller, Westphal & Reigart, 1981; Patil, Szolovits & Schwartz, 1981, 1982; Szolovits, 1979; Szolovits & Pauker, 1978).

Gorry (1970) defined diagnosis in the medical context as

> ...the problem solving activity directed toward the classification of a patient for the purpose of relating experience with past patients to him and of assessing the therapeutic and prognostic implications of his condition (p. 293).

The diagnostic model which ensues is a problem-solving approach, in which the professional's knowledge, maintained as a generalization

trom his professional education, is brought into focus in aligning the particular signs and symptoms to the closest similar known disease. The process is in three parts: the obtaining of information, the evaluation of decision alternatives, and the making of suitable diagnosis or the obtaining of additional information if the diagnosis is not yet indicated. It is a model as much of cognitive functioning as of diagnosis itself (and, perhaps suprisingly, enbodies certain strong resemblances to the model of educational diagnosis developed by Thomas). The idea of a decision tree, and a number of mathematical properties associated with such processes, have been explicated in detail (see Jacquez, 1972; Lauder, 1981); the decision tree enters the physician's strategy at the point of evaluating decision alternatives. The model is carried further by such writers as Elstein, Schulman and Sprafka (1978), who point out that many physicians do not enter the problem-solving approach without already having formed a series of working hypotheses:

> Early generation of tentative diagnostic hypotheses is ... used by clinicians to bound the regions of the potential problem space most likely to yield the solution. The subsequent workup is planned to permit testing or refinement...The method used to narrow diagnostic hypotheses and reach closure about problems or treatment alternatives is a form of means-end analysis in which specific clinical findings or clusters of finding serve as operators or movers to reduce the distance between the point where the problem solver is and where he would like to go (p. 278).

In a massive study of diagnosis and computerization in medicine, Williams (1981) presented a series of viewpoints about the diagnostic

process oriented around the orderly and logical clustering of phenomena by the observer. A major question posed by Williams is "when to study and when to act," a question which can be addressed by categorical, probabilistic, artificial intelligence, and pattern recognition models, each of which carries an extended and precise mathematical definition.

> ...Categorical approaches are particularly appropriate when the individual..."doesn't know where to start", when he seeks focus and context in a comples and ill-bounded area, and when decision choices may be optimized and then standardized according to categorical criteria. Probabilistic approaches are most useful for limited and clearly bounded problems with mathematically manageable numbers of variables. When "good" and relevant data are available, classic probabilistic approaches are applicable and may be used to support and refine expert decisions. When such data are not available, expert judgment may be codified using pseudoprobabilistic techniques and plausible reasoning, procedures that are also important in propagating even well supported uncertainty estimates, derived from classic probability, between models at different levels (vol.1, p.156).

The diagnostic situation in medicine involves, in its simplest form, the nature of the illness, the skills of the professional in discovering the exact specifications of that illness, and the tools available to aid that discovery process. In the first two areas the last decade has seen extensive research in statistical modeling of diagnostic classification, diagnostic probabilities, optimization strategies, and decision paths. In the last area, there has been an explosion of effort in relation to computerization of the diagnostic process.

A number of writers (Barr & Feigenbaum, 1982; Blois, 1980; Rogers, Ryack and Moeller, 1979; Weiss, Kulikowski, Amarel, & Safir, 1978) have provided overviews of computer-aided medical diagnosis. Over the past two decades, several extremely sophisticated interactive inquiry programs have been executed; the end user is prompted for specific information and shown, at appropriate places, the variety of possible diagnoses under consideration. MYCIN, for instance, utilizes a strategy of narrowing its options based on its conversation with the medical professional at a computer terminal until a point at which it can state a diagnosis, its confidence in that diagnosis, some alternative diagnoses if applicable, and a recommendation for course of treatment in both expected and adverse circumstances. The typical configuration of a computer-based diagnostic system involves a disease-symptom database, a combination of heuristic and statistical algorithms for developing decisions, and through the input of the medical professional, interactive contact with the target case during the diagnostic process and again upon confirmation of the diagnosis. The last step provides a feedback mechanism with which the program can validate its database. These approaches are not without controversy (see discussion section below) but the potential for computerization of the diagnostic process in medicine has been thoroughly demonstrated.

Specific illustrations can be found even in areas where the experienced clinician faces a challenge. One diagnostic problem in

newborns occurs because a wide variety of congential malformation is possible yet any single physician is likely to encounter them rarely. Computer programs now exist which allow interactive probablistic diagnostic determination to be made by a computer which accesses 224 different postnatal syndromes. Bone marrow evaluation, a pathological speciality which relies on extensive amounts of complex data, is currently being conducted on a experimental basis using a microcomputer (Wheeler, 1983). The program collects data from several sources, provides textual and graphic information to the medical professional, and concludes with a Disease Attribute Matrix Score, which combines symptoms and statistical weights to yield a tentative diagnosis or ruleout. This can be accepted or returned for revision, in which instances the user enters a series of increasingly selective queries in an attempt to further refine the working hypotheses.

Probabilistic modeling of medical decision making is another topic in current development for microcomputers (Galen, 1983; Savage, 1972), apparently with success. Over the remainder of this decade, the profession anticipates increasing reliance on computer technology not only in the making of specific diagnoses to fit specific individual cases, but in enabling the medical professional to improve the entire diagnostic process.

IV. A COMPREHENSIVE MODEL OF DIAGNOSIS IN EDUCATION

A review of the successes of diagnostic theory and practice in medicine from the viewpoint of education illuminates the following

general problems. Unlike medicine, which draws from extensive experience with most disease entities, educational diagnosis seldom has the same unambiguous reference base. While medical diagnosis successfully employs probabilistic methods, educational diagnosis only occasionally has sufficient amounts of information to support probabilistic techniques. Medical diagnosis builds on strong inference, but educational diagnosis has developed only portions of the necessary inference techniques which would allow the same degree of success.

As Hennesey (1981) illustrates, educational diagnostic specialists have accumulated "a vast amount of rich data and insight to support their practices" (p. 56). Yet the present status of models of diagnosis in education is significantly behind that of diagnostic models in medicine in at least three respects. What appears to be lacking in education is the following:

> a) design of strategies: an explanation of what the diagnostic process specifically attends to (and what it ignores) as well as what it requires the professional to do and the range of optios available for doing such;
>
> b) accumulation of evidence: a definition of what constitutes sufficient information for finalizing a diagnosis and a recognition of the strengths and weaknesses of differing information-gathering strategies; and
>
> c) computerization: use of computers to aid the teacher in collecting and evaluating data towards concluding in a diagnosis.

The first two requirements deal with the scope of the diagnostic inquiry. Thomas' first "critical question": what are the specific objectives which the individual student is expected to have achieved?

The appropriate signs and symptoms are those which point to some failure in expected achievement with those specific objectives; the working hypotheses concern the variety of plausible explanations for such a deficit. At that point, the second requirement indicates that the next task is to discover data which will narrow the list of working hypotheses appropriately.

Within this context, a generalized model, adapted from Burke (in Williams, 1981) with permission, shows how the task of diagnosis fits between the problem and the management solution. Figure 3 traces the steps of this generalized model of diagnostic process. Initial signs and symptoms are organized, following a theoretical base if possible, such that an initial profile of the student's weaknesses can be drawn together. This profile needs to address the target deficit with sufficient specificity (the substance of the area of achievement must be represented adequately) and with sufficient selectivity (the range of performance within the area of achievement must bracket the child's present capabilities) (Weiss, 1983). Ample consideration must also be paid to instructional history (Tatsuoka & Birenbaum, 1979). Working hypotheses are developed, the more formally associated with theory the better, based on an initial understanding of the pattern of responses, and from these hypotheses the most germaine diagnostic test strategies (elaborated in the following section) are brought into play. "Pattern," in the context of individualized diagnostic assessment, is used to reflect unusual responses to a set of test items, or a set of

specific erroneous responses acros similar items in a test. (For some testing strategies which explore the latter, see the accompanying paper by Choppin, 1983.)

Following the development of initial hypotheses, the ideal construction of a diagnostic process stems from the professional's careful reading of the evidence to date and sequencing of steps to gather additional evidence, until one of three actions can occur:

a) the initial hypotheses concerning the specific educational problem are supported by the tests;

b) the initial hypotheses are supported but with and unacceptable level of ambiguity;

c) the initial hypotheses are excluded.

If the initial hypotheses are supported by the tests, no further testing is required and the examiner moves, with some certainty, to the task of implementing an appropriate remediation, tailoring of the curriculum, re-education or referral. The examiner arrives at the diagnostic end point with confidence and can optimize the selection of a management strategy for the use. However, the initial hypotheses may not be completely supported by the tests; further testing which might lend clarity may be too costly in time or money. With some degree of uncertainty the examiner moves to the management of the case (and such management may consist of a referral for more specialized testing or simply waiting for some favorable turn of events). While the examiner traces the same path on Figure 3, as for the successful

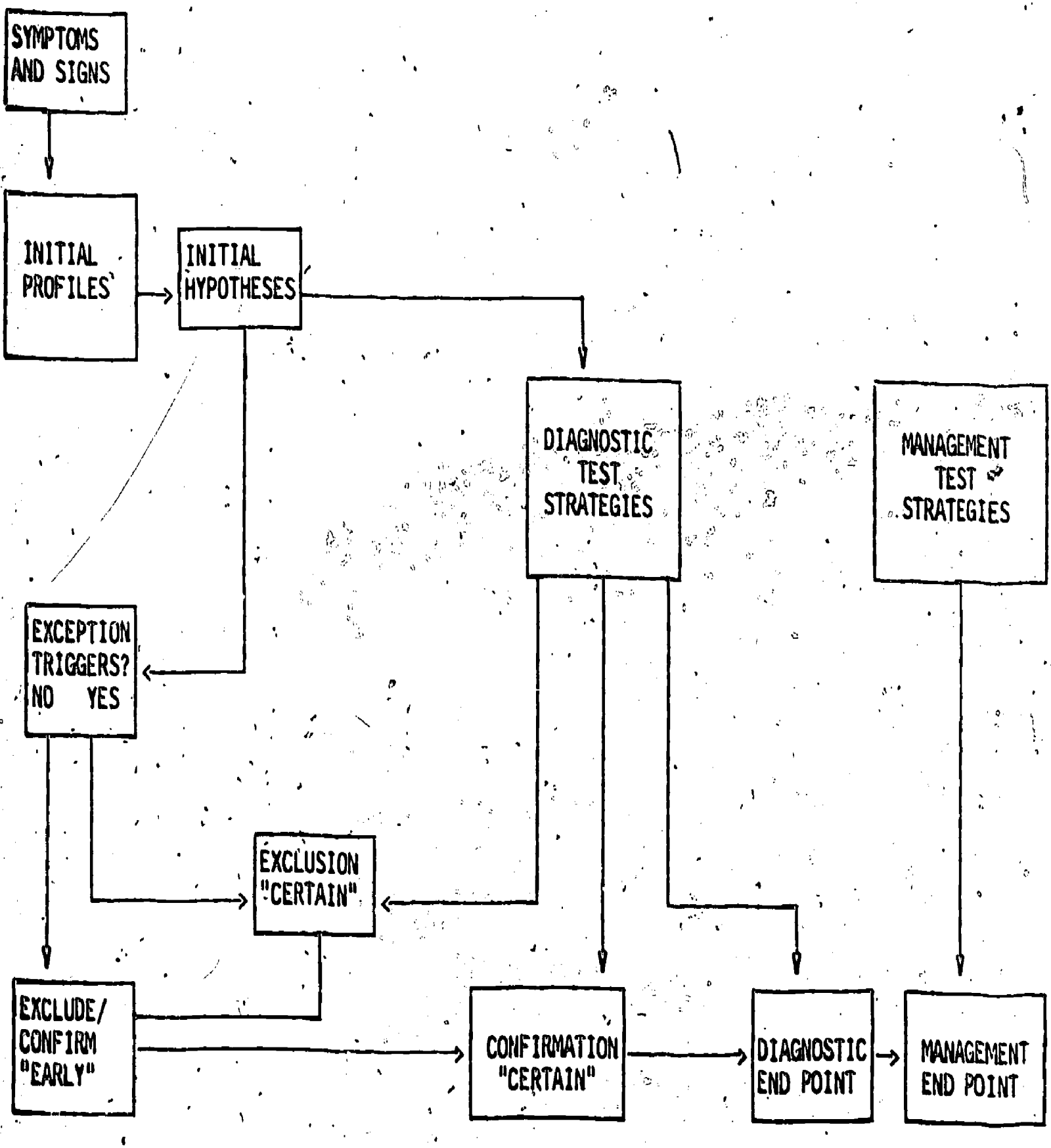


FIGURE 3. Complex diagnostic and management processes. (From Burke, M. D., Mount Sinai Hospital, Minneapolis, Minn. With permission.)

diagnosis, the outcome is expressed in less certain items. Alternatively, the examiner can use further testing to investigate whether the inital hypotheses can be overturned and the working diagnostic interpretation excluded ("exclusion certain"), or whether a different approach to the problem can generate confirmation of the inital hypotheses from separate perspective ("confirmation certain". the initial hypotheses are excluded by one of three approaches. The diagnostic testing may prove them untenable. Some "early" exclusion criterion such as strong evidence from prior testing or another professional, is provided which obviates the need to explore the initial diagnostic hypotheses further. Or those hypotheses may be excluded by an "exception trigger," a critical finding that manifests itself in psychological or educational difficulties but stems from a completely different domain altogether, for example, organic illness. These exclusions all lead the examiner away from the diagnostic endpoint in the lower right corner of Figure 3, and each implies that the inital hypothese were unsatisfactory. Further work is required, not likely involving a second look at the initial profiles of educational problems to generate a new act of working hypotheses.

At this stage, the model has served to alert the teacher to the possibility that a) initial hypotheses fit within a context of both available evidence and theory, and b) these working hypotheses help determine both what further evidence to gather and what exceptions to consider at the same time. More detail about the operations within

diagnostic testing is offered below; at this juncture, however, it is important to note that three goals in tracing diagnosis in Figure 3, from top left to lower right are to do so quickly, efficiently, and with a high level of confidence. These three are not entirely exclusive but practical considerations mitigate heavily against the professional proceeding well on all three accounts unless the data are also of high quality.

Acquiring data to support (or remove) a working hypothesis can proceed in several ways. Thomas (1983) supplies six possible sources of data: standardized tests, teacher-made tests, worksheets and regular student assignments, unrecorded observations, rating scales and interviews. The section which follows explores options for formal test strategies in diagnosis. The present state of educational testing in diagnosis is just emerging from an exclusive reliance on conventional ad seriatim testing and moving into ich variety of other strategies, some of which are set forth in Figure 4. The figure portrays schematically the movement made by the student when faced with a single test item and the ensuing possible decision points available to the examiner in each of four test strategies. The four general schemes are:

> a) ad seriatim testing -- test items are administered from first to last. No change in sequence is contemplated during the test, and, generally the evaluation of the diagnostic hypothesis is not begun until completion of the test. Most conventional educational testing and a majority of existing tests designed to be inherently diagnostic in application proceed in this manner.

b) answer-until-correct testing -- test items are administered from first to last, but a wrong answer returns the student to another opportunity to respond to the same item again, with available answers reduced by one. The evaluation of diagnostic hypotheses occurs as the student repeats the same answer strategy and obtains similar sequences of wrong answers from item to item.

c) compress-decompress (or "stradaptive" [Thompson & Weiss, 1980]) testing test -- items are administered according to a selection rule or structural lattice which allows a correct response to one item to lead to an item of greater complexity, while a wrong response to the first item leads next to an item of greater simplicity. The evaluation of diagnostic hypotheses occurs as the student repeatedly selects similar erroneous responses across items, and/or selects correctly at one level of test complexity but not at the next higher level; and/or selects dissimilar responses across items of the same complexity.

d) developmental testing -- test items are adminstered serially, often across an extended period of time. The student's response to each item is codified in multiple ways, which may include appraisals of the method or methods the student utilized to reach an answer, the type of answer given, how the student chooses to represent that answer in some formal way such as with text or symbols, and/or how the student reconstructs the original problem from the representation she made earlier. Evaluation of diagnostic hypotheses is possible upon complete codification of scores to each item.

Each of the four "maps" for traveling through a test has been used for tests which are not inherently diagnostic in nature. Nor do the four provide either an exhaustive review of all possible test design strategies nor necessarily a set of practically exclusive heuristics: it is entirely possible that advantages of one or another of the designs can be folded into a combined form of testing, and/or that a single test could begin with one scheme but branch to another at some decision-point. However, the primary reason for

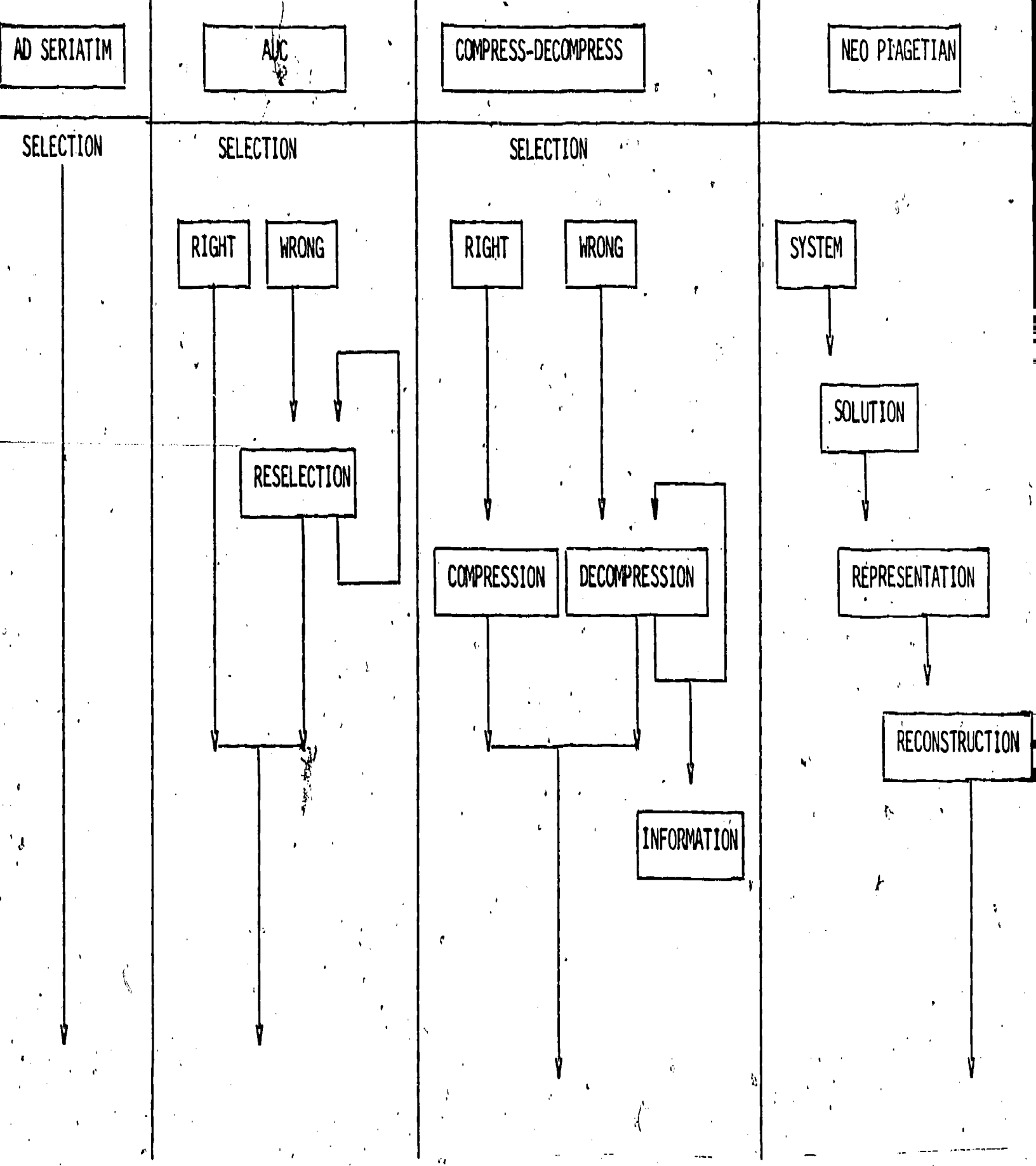


Figure 4. Diagnostic test strategies: four heuristics

distinguishing from "maps" at persent is to demonstrate the differing sources of diagnostic information that occur within each:

Ad seriatim testing: diagnostically useful information is available at the end of a complete sequence of items, but only under special circumstances is information available before testing is terminated.

Answer-until correct testing: diagnostically useful information is available whenever students select incorrect answers, and such information can be used to terminate testing before the test item. However, the information provides no immediate guidance as to sources of error.

Compress-decompress testing: diagnostically useful information is available after each student response, because the correctness of the respose is used to determine the next item to be presented. The nature of the error made if the response is incorrect can be evaluated. The student may work towards some "balance-point" within a domain, in which more difficult items cannot be answered without error while less difficult items pose no problem.

Neo Piagetian testing: diagnostically useful information is available while the student is making a response, after the student has completed the response, as the student works to draw or write down the problem as a representative of his/her thinking, and as the student views that drawing or written narrative and talks about his/her memory of the problem and the response.

All four approaches have analogues in medical diagnosis. The first, seriatim testing, reflects the protocol followed in obtaining a patient's responses to a standard family medical history. The patient goes straight through until the final item without interference from the medical professional. The second, answer-until-correct testing, mirrors the protocol used when portions of that same history are readministered orally for purposes of confirmation or further detail. The fourth, developmental testing follows to some degree the multi-modality testing used in such complex arenas as neuropathological diagnosis, in which the professional uses a wide range of dissimilar tests over a period of time in order to isolate a specific impairment.

The third approach, compress-decompress testing, reflects the more complex protocols frequently required to diagnose those problems for which multiple alternative explanations are not easy to rule out. As the professional begins to believe s/he has acquired information which fits, that information is incorporated (or "compressed") into a more encompassing understanding of the problem, until, at some point in time, sufficient confirmatory data is in hand to allow, without further delay, a diagnosis and a plan of medical care. However, as the professional gathers information which is disconfirmatory, the diagnostic process now moves to "decompress" the available information, and if necessary gather even more data, until a plausible alternative hypothesis emerges with some degree of certainty.

The general model of diagnostic process allows a perspective on possible computerization. First, using a computer to accomplish this process requires that enough is already known about particular sets of errors or problems to facilitate the formation of initial hypotheses. If true, then each of the four heuristic designs of Figure 4 can be brought within the strictures of the real-time interactive computer. Second, with the computer used for both administration and statistical analysis, the teacher can engage interactively, during test administration or after, to provide additional information for a categorical or probabilistic diagnostic assessment predicated on solid mathematical principles (Bock & Mislevy, 1982; Tatsuoka & Linn, 1983; Weiss, 1982). Further comments about computerization follow later in this paper.

DISCUSSION

1. Diagnostic interpretations of illustrative data

The following is a brief summary of findings from four studies of test performance and diagnosis: an ad-seriatim test of language arts skills (presented in detail as a separate report), an answer-until-correct test of arithmetic skills, a compress-decompress prototype test of understanding of science, and a developmental test of elementary number concepts. The first, second, and fourth tests adhere closely to the first, second, and fourth heuristics of diagnostic testing presented earlier (ad seriatim, answer-until-correct, and neo Piagetian; the third served to

illustrate certain aspects of the third heuristic (compress-decompress) although it was delivered to students serially. The four tests were each designed to reflect very specific subject domains and were administered in different ways to different examinees:

Language arts: a 92 item test of pronoun understanding, in which the development of the items followed a rigorous structural interpretation of pronoun usage and complexity, and of the sentence context within which target pronouns were embedded. The test items were developed to reflect the application of six rules of grammar in usage of first person plural, third person singular, and third person plural constructions. For each rule, six items required the examinee to recognize and select the correct form and rule without making inferences, and six required the student to infer the correct form and concept from the item stem. This test was administered as a paper-and-pencil test to 49 Fluent English Proficient and 79 Limited English Proficient sixth graders in Los Angeles County.

Arithmetic skills: A 10 item test of arithmetic skills involving addition, subtraction, and multiplication approximately geared to the sixth grade level. This test was administered on a one-to-one basis by microcomputer following the answer-until-correct strategy (items were presented again if the examinee's response was wrong, until such time as the right answer was selected from the remaining options or itself was the only remaining option). Examinees for this test were 68 fourth through eighth grade students attending summer courses in computing at UCLA.

Understanding of science: a 20-item test of selected concepts in science, constructed with attention to two key factors: rational construction of distractors within each item and between related items, and hierarchical ordering of items by complexity. The test involved three kinds of distractors: logical fallacy, intuition distraction, content distraction, presented in items of low, medium and high difficulty in form selected topics in science. This test was administered as a paper-and-pencil test to 190 students representing a very large range of exposure to science concepts: high-talent private junior high students, a mixed range of

ability levels in public high school classes, and entering college freshmen studying Introductory Biology, all in Los Angeles County.

Elementary number concepts: a multi-part developmental test of selected concepts in counting and constructing one-and two-digit numbering wooden blocks. The test examined concepts of numbers including counting, adding/subtracting, constructing with modular blocks and constructing combinations. This test, building extensively on neo-Piagetian theory, was administered adaptively on a one-to-one basis by trained examiners to 99 kindergarten through second grade pupils in Santa Barbara County.

The language arts test data was extensively analyzed by methods which address group and subgroup distinctions and differences between facets of the item design. Of interest to the present report are those findings which address individual performance. What emerges is a profile of each examinee's performance presented as proportion of correct response to the item facets, annotated by a statistic which addresses correspondence between profiles and various generalizability coefficients at each facet. Selected cases show substantial variation in relation to the item facets, but classical test strategies show that the test itself is reliable and that certain expected patterns of performance (lower success with context-embedded pronouns than with the same pronoun in an item without the embedding, for example) generally hold true. Diagnoses of individual problems with particular forms of pronoun usage can be easily drawn from examination of patterns of performance on a case-by-case basis. In this sense, the predominant meaning of "pattern" is a profile of subscale scores. Some difficulties were common to all students, and thus not inherently

diagnostic. Other difficulties appled to selected students, and for these the individual profiles should yield diagnostically useful information. Profile for students with fluency in English were paralleled by the profiles for Limited English Proficient students. Further discussion is found in the accompanying report (Webb et al, 1983).

The arithmetic skills answer-until-correct test was analyzed by a variety of methods which primarily address issues of reliability and selection (for an extended discussion see the accompanying reports by Wilcox, 1983). The procedures evaluate the probability of correctly determining that a given examinee knows a given item. Using the answer-until-correct model, probabilities were estimated for each item: the first six had probabilities exceeding .85; the remainder were at least .71 or greater. The probability of at least seven correct decisions (i.e., whether it was correctly determined that an examinee knows an item) could also be estimated for this test. Using recent psychometric developments, it was determined that an estimated lower bound of this probability value was .70, while the estimated lower bound of at least six out of ten correct decision was .83. Thus the test appears to be fairly accurate, although additional scoring rules which are useful in improving accuracy had minimal effect with this dataset.

Essentially, this analysis is premised on a latent-trait model of examinee behavior, in which the harder items generate more inaccurate

measures of whether or not the student knows the correct answer, and at the same time call into play more guessing behaviors. The probability of making a given number of correct decisions given the total number of items is analogous to a test of reliability, in that both generate a single number which characterizes the adequacy of the set of test items. These procedures, however, do not specifically speak to the problem of individual diagnosis. Instead, that issue can be taken up by other measures of individual performance using the probabilistic information of correct determination of the individual's latent state as a base. However, it should be noted that methods which using the first response only as indication of right or wrong will not comport with the answer-until-correct analyses, because the latter are able to take the full nature of the response behaviors to a given item into account. The only case in which traditional measure of 1/0 scoring based on first response only will agree with answer-until-correct analyses is the impossible case in which examinees never find the correct answer if they miss an item on the first try.

It is important to note that answer-until-correct testing utilizes a highly specific definition of "pattern" in analyzing test performance: "pattern" is taken to mean repeated attempts to secure a correct answer, with both the number of such attempts within a given item and the number of items requiring repeated attempts having a direct impact on the associated statistics. The use of

answer-until-correct testing to diagnose individual difficulties in a content domain could rely upon these two elements within the individual pattern of performance.

Research efforts by a diverse group of educational and psychological workers have explored the nature of logical thinking and hierarchical structuring of knowledge (ct. Cotton, Gallagher and Marshal, 1977; Dreyfus and Jungwirth, 1980; Rodgin, 1955) but in general there remains a great deal of disagreement as to how hierachies may be assembled and to their validity and repeatability even within narrowly defined topic areas. The point of view adopted is critical in determining the rest of the research that ensues. In the area of structure of mathematical concepts in school children, for example, recent publications by workers in England (Hart, 1981; Osborne, 1983), Russia (Krutetskii, 1976) and Finland (Keranto, 1981) appear to share very little in common. Despite this, work has progressed towards analyzing tests in selected topics diagnostically. In the area of diagnostic testing of mathematical abilities, Birenbaum and Tatsuoka's (1980) contribution is but one of a series from workers at the University of Illinois; for diagnostic testing of science concepts several studies can be cited which have proved at least partially successful (cf. Bartov, 1978; Gorodetsky & Hoz, 1980; Long, Okey & Yearny, 1978). Johnstones' (1981) review provides an excellent overview of problems in diagnostic testing in science.

The compress-decompress prototype test of understanding of science was an attempt to incorporate structural hierarchies relating to conceptual understanding of selected science topics with a rule-based algorithm for the construction of each distractor to each item. A three level comprehension strategy (factual knowledge, recognition of principle as well as factual knowledge, and application as well as recognition of principle) was used to construct a twenty item test. Each item's four choices were restricted to a logical fallacy, an intuition, a faulty content similarity, and the correct response. (A detailed report of the results of this study is found in Shaha (1983)).

In the context of the present paper, the important data elements from this endeavor are three in number: first, the general profile of responses across correct and incorrect alternatives for related items at different levels of comprehension; second, the general profile of responses for those items across the same comprehension level; and third, the degree of variation of performance of individual examinees with regard to both related items and like levels. Diagnostic interpretations can be derived directly from the third "pattern" listed here: the word "pattern" is taken to refer both to specific sorts of erroneous responses and to consistent (or inconsistent) indicators of conceptual level across differing subtopics. In this test, missing more than one item at any level of comprehension was almost always matched by a mass of at least two more difficult items

in the same domain. While error patterns did not appear consistant across the entire test, there were consistent patterns of error, particularly logical fallacy and intuition errors, within topics.

Developmental testing of children's number concepts among kindergarten, first and second graders was carried out in a variety of separate subtopics on a one-to-one basis by trained examiners; this extended dataset has been kindly supplied by Dr. Jules Zimmer of the University of California Santa Barbara. The data consist of four separate appraisals by the examiner for every target response: the strategy by which a given number concept problem was solved, the accuracy of that solution, the ability of the child to draw a version of what she or he did to handle the problem, and the reconstruction of the solution from that drawing a week later. Each of four sets of problems was evaluated in this manner, yielding extensive data which could be characterized as follows for the majority of cases:

- the accuracy of the response was usually related to the strength of the strategy employed by the child.
- the ability to represent the problem was usually related to the accuracy of the solution.
- the ability to reconstruct the problem from the representation was often related to initial strategy.

The diagnostic portion of the study concerned the question as to whether patterns of performance by a minority of students were erratic over the sets of problems. Here the use of the word "pattern" is taken to mean inconsistent behaviors across differing situations.

Consistent behaviors were taken as indication of level of mathematic ability, while inconsistent behaviors were viewed as the key to how children would face specific trouble spots in their own understanding of number concepts. Because of the close contact between students and teacher at this school level, this dataset represents the one present source for which statistical flags for individuals can be compared blindly to the informal assessment made by their teachers. In brief, the diagnostic question was approached by evaluating the extent of intrasubject agreement across problem sets. Within the subtopics, those individuals who were substantially inconsistent overall were flagged and the number of such flags totaled. Seven students were identified as having a pattern of responses which revealed erratic performance. These seven, plus two others, were the same students independently seen by teachers as currently in educational difficulty or likely to require close attention during the present school year. Further detail will appear in forthcoming reports.

2. Advantages and disadvantages of diagnostic testing in education.

It is inaccurate to paint too rosy a picture of computerized diagnostic testing in education at this time. Despite extensive psychometric research, the primary restrictions revolve around the relatively course grain of measurement in educational testing. That is, for any single test response or collection of test behaviors in most areas of education, no responsible party claims to know the complete underlying cause or causes. In computerized psychological

testing, in contrast, certain test responses, appearing as a set, can be linked with very high confidence to a narrowly defined, and thus diagnostically strong, set of likely explanations. Likewise, in many instances in computerized medical diagnostic testing, once beyond a critical mass of evidence there are few other plausible outcomes of a testing algorithm in addition to the one or two primary diagnoses.

The obvious success with which diagnosis takes place in the field of medicine cannot be matched by comparable successes in most of the field of education. A variety of interrelated explanations for this current state of affairs are available, among which are problems of diagnostic definition, test construction, and practical management.

However, once a certain number of problems are favorably resolved, it appears that using computers to score and interpret diagnostic tests in educational settings can accrue the same advantages as in the current practice of computerized testing in psychology and medicine. First is the significant accumulation of hard evidence in the form of a computer databank of diagnostic indicators. Until computerization, this bank exists mainly as personal experience. Until computerization, use of logically rigorous diagnostic procedures is markedly limited by being tied to paper-and-pencil instruments. Until computerization, adaptive exploration of possible diagnostic pathways is limited by the patience and agility of the teacher in bringing various parts of a test instrument to the examinee at the appropriate moments.

With an appropriate backlog of data, a computer-driven scoring procedure can efficiently sort results of a test administration following algorithms regarding hypothesis likelihood. The procedure can evaluate an extended range of findings cooperatively across several different tests of the same individual. The procedure can explore competing alternatives without prejudice, delivering in conclusion a summary of findings, a statement as to the confidence level of those findings within the context of the given tests, and potentially useful avenues for student remediation.

One key problem requiring further research is the problem of properly encapsulating any respectable cross-section of subject matter within the highly restricted rules which govern both diagnostic testing and computerization. That is, even the most flexible diagnostic strategy, managed by the most intelligent and "user-friendly" computer programs, is likely to involve severe trade-offs between optimal measurement characteristics, the available level of "understanding" of language built into the program, and practical issues of both test applicability and diagnostic interpretation. Experience with a promising computer-driven educational diagnostic algorithm in the Netherlands (Gobits, 1973) validates these concerns:

> One can expect...severe difficulties...when trying to convey meaning by a language of very restricted code, i.e. a language with severe regulations as to how the form should be. In fact it turned out to be practically impossible to shape richer subject matter content into the highly regulated forms of the suggested

> 'language'... The moral, of course, is that any 'language' one devises for testing and feedback purposes with more restricted code then natural language will pose practical problems...and take additional instruction. (R. Gobits, personal communication, 1983).

Another problem to be resolved is the closer integration of test objectives with curricula. This requirement is addressed infrequently but must be stressed. Even the most elegant statistically based computer-managed test sequence comes to naught if not tied to the curriculum. The relative success of diagnostic testing in reading and simple arithmetic may rest on the extensive acceptance in most school systems of reading and arithmetic curricula which generally cover the same explicit goals at the same grade levels even when teaching methods differ widely. However, many subject domains within American primary and secondary education, such as the physical and biological sciences, topics in mathematics beyond elementary algebra, and computing, are treated uniquely even between neighboring schools in the same district. With little common ground to stand on diagnostic testing may be much more difficult to organize on a broad scale.

However, it is only fair to indicate that many of the concerns which pertain to educational diagnosis and computerization exist in the best of efforts involving artificial intelligence to solve diagnostic problems in medicine. Szolovits and Pauker (1978) evaluating a series of computerized medical diagnosis programs, list several important shortcomings:

> 1. Programs which deal with relatively broad domains...have inadequate criteria for deciding when a diagnosis is complete...The programs continue exploring less and less sensible additional hypotheses...
>
> 2. Because the initial strategy...is to use every significant new finding...and because this strategy remains through the programs' operation, new hypotheses are continually being activated...
>
> 3. Part of the routine developed by clinicians is an appropriate order for acquiring information systematically. Computer diagnosticians tend either to enforce such an order too strictly...or not at all...
>
> 4. The programs rely on a global assessment scheme, but they use too weak semantics for the states over which they try to compute approximate probabilities...None of the programs can dynamically distinguish among...aggregate hypotheses... Yet there are therapeutic and strategic decisions which hinge on just such distinctions... (pp139-140)

Advances in computerized medical diagnosis since publication of this important article have attended to, but have yet entirely resolved, these concerns.

Diagnostic clarity is lacking in general educational practice, the areas of reading and speech aside, partially because, unlike medicine or psychology, the field of education has only occasional databases which go beyond summary scores by which to examine one or more normative patterns of skill acquisition. Moreover, the processes of skill learning even within very restricted areas such as arithmetic are only beginning to be understood at the same level of detail as, for example, acquisition of object permanence in infants. In speech and reading diagnosis, and to some degree in elementary operations in

arithmetic, diagnostic instruments are available which allow efficient strategy, interpretation, and management. This success stems in part from a long cumulative history of effort in these areas and in part from the ability to define very closely the exact skills to be targetted at each step of the student's development. However, even in speech, reading, and arithmetic, the field labors under an excessive number of plausible competing hypotheses, many of which compound one another. Thus the task of obtaining clear and unambiguous diagnoses is seldom one which can be completed with a large degree of confidence.

Test construction has advanced in countless respects during the last decade, including in particular the mathematical and statistical developments necessary to support alternative test strategies. However, to construct an adequate diagnostic test requires an additional series of considerations: given appropriately specific definitions, can one write items, for a conventional or non conventional diagnostic test, which are jointly corroborating, exhaustive of the viable alternatives, and parsimonious? The obvious goal is to obtain reliable items which demonstrates differential prediction of future performance. Within a test the related items must be structurally coherent both in respect to item content and type of response. Yet the same items must also allow the student to give any significant logically interpretable response whether correct or erroneous.

From the viewpoint of practice, the management of a diagnostic test obviously requires more than the usual attention from the teacher, and more effort to interpret. The student, though, may treat the experience in much the same manner as any other test, including obtaining correct answers by erroneous methods and accidental guessing of correct answers (both of which make educational diagnosis especially difficult). The student may simply be sloppy in responding but the diagnostic protocol will attempt to treat every answer, right or wrong, as equally legitimate.

REFERENCES

Anderson, C.J. The use of the Woody Scale for diagnostic purposes. Elementary School Journal, 1918, 16, 770-781.

Arter, J., & Jenkins, J. Differential diagnosis - prescriptive teaching: A critical appraisal. Review of Educational Research, 1979, 49, 517-555.

Barr, A. & Feigenbaum, E.A. The Handbook of Artificial Intelligence. Los Altos, California: William Kaufman, 1982.

Bartov, H. Can students be taught to distinguish between teleological and causal explanations? Journal of Research in Science Teaching, 1978, 15, 567-572.

Bergan, J.R., Towstopiat, O., Cancelli, A.A., & Karp, C. Replacement and component rules in hierarchically ordered mathematics rule learning tasks. Journal of Educational Psychology, 1982, 74, 39-50.

Birenbaum, M. Error analysis - it does make a difference. Doctoral Dissertation, University of Illinois at Urbana-Champaign, 1981.

Birenbaum, M., & Tatsuoka, K.K. The use of information from wrong responses in measuring students achievement. (Research Report 80-1). University of Illinois, Computer-based Education Research Laboratory, 1980.

Birenbaum, M., & Tatsuoka, K.K. The effect of a scoring system based on the algorithm underlying the student's response patterns on the dimensionality of achievement test data of the problem solving type. Journal of Educational Measurement, 1983, 20, 17-26.

Blois, M.S. Clinical judgment and computers. New England Journal of Medicine, 1980, 303, 192-197.

Bock, R.D., & Mislevy, R.J. Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 1982, 6, 431-444.

Box, G.E.P., & Tiao, G.C. Bayesian inference in statistical analysis. Reading MA, Addison-Wesley, 1973.

Brown, J.S., & Burton, R.R. Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 1978, 2, 155-192.

Connolly, A.J., Nachtman, W., & Pritchett, E.M. KeyMath diagnostic arithmetic test. Circle Prince, Minn: American Guidance Service, 1971.

Cotton, J.W., Gallagher, J.P., & Marshall, S.P. The identification and decomposition of hierarchial tasks. American Educational Research Journal, 1977, 14, 189-212.

Delandsheere, G. Diagnostic assessment procedure. In T. Husen and T.N. Postlethwaite (Eds.). International encyclopedia of education. Oxford: Pergamon Press, in press.

Dreyfus, A., & Jungwirth, E. A comparison of the 'prompting effect' of out-of-school with that of in-school contexts on certain aspects of critical thinking. European Journal of Science Education, 1980, 2, 301-310.

DeGroot, M.H. Optimal statistical decisions. New York, McGraw-Hill, 1970.

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. Medical problem solving, an analysis of clinical reasoning. Cambridge, Massachusetts: Harvard University Press, 1978.

Fortes, M. A study of cognitive error. British Journal of Educational Psychology, 1932, 2, 297-318.

Furlong, F. & Miller, W. Diagnose: computer-based reporting of criterion referenced test results. Educational Technology, 1978, 8, 37-39.

Galen, R.S. What can microcomputers do for medical decision making. Diagnostic Medicine, 1983, 6, 75-78.

Gheorghe, A.V., Bali, H.N., Hill, W.J., & Carson, E.R. Dynamic decision models for clinical diagnosis. International Journal of Bio-Medical Computing, 1976, 7, 82-92.

Gobits, R. Diagnosis in a computer managed instructional system, in H.F. Crombag and D.N. DeGruijter, Contemporary issues in educational testing. The Hague, Netherlands: Mouton, 1972 (abstract).

Gorodetsky, M., & Hoz, R. Use of concept profile analysis to identify difficulties in solving science problems. Science Education, 1980, 64, 671-678.

Gorry, G.A. Modeling the diagnostic process. Journal of Medical Education, 1970, 45, 293-302.

Green, B.F. Adaptive testing by computer, In R.B. Ekstrom (Ed.). Measurement, technology and individuality in education. San Francisco: Jossey-Bass, 1983.

Hart, K. (Ed.). Children's understanding of mathematics 11-16. London: John Murray, 1981.

Hennessy, J.J. Clinical and diagnostic assessment of children's abilities: traditional and innovative methods. In P. Merrifield (Ed.), Measuring human abilities. San Francisco: Jossey-Bass, 1981.

Hunter, M. Diagnostic teaching. The Elementary School Journal, 1979, 80, 41-46.

Jacquez, J.A. (Ed.). Computer diagnosis and diagnostic methods. Springfield, Illinois: C.C. Thomas, 1972.

Johnstone, A.H. Diagnostic testing in science. In A. Levy and D. Nevo (Eds.), Evaluation roles in education. London: Gordon and Breach, 1981.

Kallom, A.W. The importance of diagnosis in educational measurement. Journal of Educational Psychology, 1919, 10, 1-12.

Keranto, T. Lukukasitteen kehittyminen ja kehittaminen: Matemaatiis-loogiset perusteet ja luvun kognitiivinen rakentuminen. (Acquisition and aided development of the concept of number: Mathematical-logical principles and the cognitive structuring of number). Acta Universitatis Tamperensis, Ser A, (Vol 125). Tampere, Finland: 1981.

Krutetskii, V.A. The psychology of mathematical abilities in school children. Chicago: University of Chicago Press, 1976.

Lauder, I.J. Latent variable models for statistical diagnosis. Biometrika, 1981, 68, 365-372.

Long, J.C., Okey, J.R., & Yeany, R.H. The effects of diagnosis with teacher-or-student-directed remediation on science achievement and attitudes. Journal of Research in Science Teaching, 1978, 15, 505-511.

Miller, M.C., Westphal, M.C., & Reigart, J.R. Mathematical models in medical diagnosis. New York: Praeger, 1981.

Nesbit, M.Y. The Child Program: Compute Help in Learning Diagnosis of Arithmetic Score (Curriculum bulletin 7-E-B). Miami: Dade County Board of Public Instruction, 1966.

Novick, M.R. Bayesian considerations in educational information systems. Iowa City, Iowa: American College Testing Program, ACT Research Report #38, 1970.

Novick, M.R., Jackson, P.H., Thayer, D.T., & Cole, D.S. Application of Bayesian methods to the prediction of educational performance. Iowa City, Iowa: American College Testing Program ACT, Research Report #42, 1971.

Okey, J.R. Diagnostic testing pays off. Science Teacher, 1976, 3, 27.

Osborn, H.H. The assessment of mathematical abilities. Educational Research, 1983, 25, 28-40.

Patil, R.S., Szolovits, P., and Schwartz, W.B. Casual understanding of patient illness in medical diagnosis. Proceedings of the IJCAI, 1981, 893-899.

Patil, R.S., Szolovits, P. & Schwartz, W.B. Information acquisition in diagnosis. Proceedings of the AAAI, 1982, 345-348.

Paulu, E.M. Diagnostic testing and remedial teaching. Boston: Heath, 1924.

Rodgin, D.W. A factor analytic study of fallacies in logical thinking. Unpublished Doctoral dissertation, Purdue University, 1955.

Rogers, W., Ryack, S.G., & Moeller, G. Computer-aided medical diagnosis: literature review. International Journal of Bio-Medical Computing, 1979, 10, 267-289.

Savage, L.J. Diagnosis and the Bayesian viewpoint - In J.A. Jacquez (Ed.), Computer diagnosis and diagnostic methods. Springfield, Illinois: G.C. Thomas, 1972.

Smith, R.M. (Ed.). Teacher diagnosis of educational difficulties. Columbus, Ohio: Charles Merrill, 1969.

Spearman, C. The origin of error. Journal of General Psychology, 1928, 1, 29-53.

Szolovits, P. Artificial intelligence and clinical problem solving. Cambridge, MIT Laboratory for Computer Science, 1979.

Szolovits, P., & Pauker, S.G. Categorical and probabilistic reasoning in medical diagnosis. Artificial Intelligence, 1978, 11, 115-144.

Tamir, P. & Kemp, R.F. Cognitive preference styles across three science disciplines. Science Education, 1978, 62, 143-152.

Tatsuoka, K.K., & Birenbaum, M. Danger of relying solely on diagnostic adaptive testing when prior and subsequent instructional methods are different. 1979. (ERIC Document No. ED 183 608)

Tatsuoka, K.K., Birenbaum, M., Tatsuoka, M.M., & Baillie, R. A psychometric approach to error analysis on response patterns. (Research Report 80-3). Urbana, Ill.: University of Illinois, Computer-based Education Research Laboratory, February 1980.

Tatsuoka, K.K., & Linn, R.L. Indices for detecting unusual patterns: links between two general approaches and potential applications. Applied Psychological Measurement, 1983, 7, 81-96.

Tatsuoka, K.K., & Tatsuoka, M.M. Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 1983, 20.

Thomas, R.M. Diagnosing and treating learning difficulties. Unpublished manuscript, 1983.

Thomas, R.M. A model of diagnostic evaluation. In A. Levy and D. Nevo (Eds.), Evaluation roles in education. London: Gordon and Breach, 1981.

Thompson, J.G., & Weiss, D.J. Criterion-related validity of adaptive testing strategies. 1980. ERIC Document No. ED 191 882, 1980.

Tyler, R.W., & White, S.H. Testing, Teaching and Learning. Washington, D.C.: NIE, 1979.

Uhl, W.L. The use of standardized materials in arithmetic for diagnosing pupils' methods of work. Elementary School Journal, 1917, 18, 215-218.

Weiss, D.J. Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 1982, 6, 473-492.

Weiss, D.J. New horizons in testing. Latent trial theory and computerized adaptive testing. New York: Academic Press, 1983.

Weiss, S.M., Kulikowski, C.A., Amarel, S., & Safin, A. A model-based method for computer-aided medical decision-making. Artificial Intelligence, 1978, 11, 145-172.

Wheeler, L.A. Computer as consultant in bone marrow studies. Diagnostic Medicine, 1983, 6, 13-17.

Williams, B.T. Computer aids for clinical decisions. Boca Raton, Florida: CRC Press, 1982.

Willing, M.H. The encouragement of individual instruction. Journal of Educational Research, 1920, 1, 193-198.

Appendix A

METHODOLOGY PROJECT

Deliverable - November 30, 1983

SOME STRATEGIES FOR CONSTRUCTING AND VALIDATING DIAGNOSTIC HYPOTHESES

by

Bruce H. Choppin
and
David L. McArthur
Project Directors

Grant Number
NIE-G-83-0001

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California
October, 1983

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Table of Contents

List of Figures

1. Test construction strategies

Because diagnostic testing depends critically on the strength of the test items, strategies for the development of the strongest possible items are essential. Four different strategies for writing multiple choice items and their distractors are considered below.

Writers of multiple choice testing exhibit significant disagreement about the role played by distractors, the incorrect alternative responses. In part this stems from the different uses to which test scores are put. In a criterion-referenced test an incorrect response directly conveys a piece of information about the individual's achievement, but in a normative test it serves only as an aid towards ranking students on a total scale. However, some of the divergence results from conflicting views of the strategies adopted by a student to answer a multiple choice item. Thus this paper continues with a consideration of two analytic approaches to analyzing patterns of erroneous responses: contingency analysis, in which attraction to similar distractors is tested categorically, and probabilistic evaluation, in which distractor attraction is tested by a process of probabilistic differentiation among competing hypotheses.

Strategy 1. The Plausibility Criterion

Using the plausibility strategy item writers construct statements that will appear reasonable to an uninformed person, but which would be judged clearly incorrect by an expert. Two items constructed in this way from a Test of the Understanding of Science are presented in Figure 1. As a rule the correct response is written first, and then

distractor statements are constructed to match it as nearly as possible in terms of length and linguistic complexity. Further, the distractors should appear sufficiently plausible to individuals with low achievement, so that a substantial proportion of examinees would be inclined to choose one of them rather than the correct answer. Estimates made by item writers of the plausibility of particular distractors are prone to error, and it is rarely possible to get a reliable estimate of a particular distractor's drawing power without field testing the item in its complete form. A general guideline for test constructors who work in this fashion is that a distractor that attracts fewer than 10 percent of the erroneous responses is not doing its job adequately and should be replaced by a more plausible statement.

Figure 1: Items illustrating distractors generated according to a plausibility criterion.

6. We do experiments when we are learning science because:
   - A. Experiments are used to test ideas by experience.
   - B. Experiments enable us to learn better.
   - C. Experiments make learning more interesting.
   - D. We can show that we all get the same results.
   - E. It is important to learn to handle apparatus skillfully.

7. Why should one make a written note of all the observations make when carrying out a scientific investigation?
   - A. One might forget them, and they may turn out to be important later.
   - B. It is a good way to train powers of observation.
   - C. It trains one to think clearly and write accurately.
   - D. Good scientists always do it.
   - E. One is supposed to have a complete record of what has done.

Source: IEA Test of Understanding Science

This method of constructing multiple choice items, though very widespread, usually is of limited interest in diagnostic testing because the choice of a particular distractor seldom gives clear information about the learning problems of the individual testee.

Strategy 2. The Use of Most Frequent Errors

The "most frequent errors" strategy, in its simplest mode,

consists of giving test items in an open-ended format to samples of individuals at an appropriate level in order to determine the three or four erroneous responses that are given with the highest frequency. More pragmatically based than Strategy 1, it produces distractors that are plausible from the students' point of view, but it suffers from a major drawback in that many of the most frequent responses produced by students will be almost correct. Thus high ability students and experts may not be able to discriminate between correct and incorrect responses with a high rate of consistency, and overall test reliability may be low. If the student-generated distractors are modified by the test constructor to make them clearly incorrect then their built-in plausibility may disappear. Once again, note that distractors generated by this method are rarely intended to carry diagnostic information.

Strategy 3. Logical Error Analysis

Items can be designed such that the distractors reveal specific errors of logic and procedure. Figure 2 contains items from an arithmetic test in which the distractors have been designed to be chosen by students who make particular procedural errors. A student who transfers incorrectly between the tens and units column in item 9 might be expected to pick response (E). In a test in which all the items concentrate on a narrow domain of skills, such as integer addition or subtraction, it may be possible to infer certain diagnostic conditions from the pattern of responses to the whole test. However, many of the multiple-choice tests that use this

approach use quite different techniques for generating the distractors to different items, depending upon the content of each item concerned. Figure 3 gives an example of two such items from a science achievement test. Here the diagnostic information revealed by a single incorrect response is too unreliable (see Tatsuoka, Birenbaum, Tatsuoka, and Baillie, 1981) to be interpreted, and since the evidence provided by different items bears on different issues, aggregation of the diagnostic information from the items is difficult.

Figure 2: Items constructed according to a logical error analysis.

9.	53 - 26 ?	A. 33 B. 37 C. 27 D. 79 E. 47
10.	44 - 16 ?	A. 32 B. 38 C. 48 D. 28 E. 60

Figure 3: Items illustrating logical error analysis for distractors but on unrelated content.

12. Flour is a fine powder obtained by grinding wheat or other cereal grains. A pile of grain burns only very slowly whereas flour dust suspended in air is explosive? Which of the following is the best explanation of this?

A. The heat produced when small particles burn is greater than the heat produced by the burning of large particles of the same substance.
B. Grinding the grain changes its chemical composition.
C. For the same quantity of the material, small particles have a greater surface area in contact with air than large particles.
D. Small particles possess more energy than large particles.
E. The flour burns completely whereas the pile of grain does not.

13. Two given elements combine to form a poisonous compound. Which of the following conclusions about the properties of these two elements can be drawn from this information?

A. Both elements are certainly poisonous.
B. At least one element is certainly poisonous.
C. One element is poisonous, the other is not.
D. Neither element is poisonous.
E. Neither element need be poisonous.

Source: IEA Science Test 4B

Strategy 4. Theory Based Distractors

Items can be defined following a theory of the consistent parts of erroneous understanding; this strategy for item writing is comparatively rare. In a typical example, the test constructor attempts to use a theory of student cognitive behavior, a logical analysis of the subject area, or a personality theory in order to define a discrete number of response types, and to write distractors for each item that falls into one of these types. A good example of such a test is the Cognitive Preference Style in Science test developed by Kempa (Tamir & Kempa, 1978); a sample item from this test is given in Figure 4. Four styles of cognitive preference between which the test is designed to discriminate are recall, principles, questioning, and application. The item shown has no incorrect answers; instead it is hypothesized that a student whose preference is for the recall style would be most likely to select response (A), whereas a student whose preference is for application would tend to select option (B), etc. Such tests typically used not for routine assessment or diagnosis but for research, and in many cases the evidence for their theoretical validity is not strong. However, initial successes in using strict theoretical frameworks to construct such instruments suggests that it is possible to apply a more structured approach to the design of distractors for regular diagnostic instruments..

Figure 4: Example of a theory-based item structure.

A gas spreads out to fill the volume of the containing vessel.
(A) Gas particles are in a state of motion.
(B) The movement of the gas molecules enables us to experience smells at a distance from their origin.
(C) The speed of movement depends on the mass of the gas molecules.
(D) The gas molecules are in a state of perpetual motion because they possess kinetic energy.

Source: Tamir and Kempa, 1978.

2. Contingency Analysis

Multiple-choice achievement test data characteristically show a fair amount of inconsistency: even the most able students sometimes select incorrect responses, for reasons that are often unclear. Less able students sometimes select correct responses to difficult problems, again for reasons that are often unclear, probably but not necessarily guessing at random even when one might hypothesize that their level of understanding would lead them to choose one particular distractor.

Tests must be composed of many items if reliability and precision are to be achieved; however, one major goal of diagnostic testing is to form reliable diagnostic judgments from the pattern of results

of a parsimonious set of items. If meaningful diagnostic interpretations are to be predicated on the choice of a particular distractor, then the item and its distractors must themselves be strong enough to sustain it; within a reasonable probability the selection of a particular distractor must reflect the examinee's state of learning and/or mislearning in the domain.

A straightforward approach to the investigation of this issue is through contingency analysis. If items are functioning as expected diagnostically, we hypothesize that an examinee who responds with error "A" to one item will also respond with error "A" to related items. It is appropriate to require, as evidence of the diagnostic validity of paired distractors, that a significantly larger number of like-error contingencies occur than would occur if responses were random.

For each item pair a table of frequencies of all possible response pairs is evaluated by simple χ^2 to show whether the pattern is random. Inclusion of a 'correct' answer to both items is sufficient to render the χ^2 value significant. An improved test calculates χ^2 for the response grid <u>after</u> eliminating the row and column corresponding to 'correct' answers. If significant, a check can be made to determine if the predicted patterns are those that occur with unusual frequency.

<u>3. Probabilistic diagnostic evaluation:</u>

a.) Theory of probabilistic differentiation

Many of the successes of diagnosis in medical practice can be

to use of probabilistic rather than strictly categorical evaluation of the available evidence at any given time. Diagnostic testing in education can be regarded as a process of probabilistic differentiation between alternative hypotheses, one of which is that the student has actually mastered the material. By constructing a test which combines a plausible set of hypotheses based on common reasons for failure with a probabilistic scheme for evaluation (Box & Tiao, 1973; De Groot, 1970; Novick, 1970; Novick, Jackson, Thayer & Cole, 1971), the diagnostic process can be made efficient.

One question which can be addressed from several angles is the question of optimal stopping: how much evidence can be regarded as sufficient? The mathematical concept of "martingale" allows one vector of information (read student's responses to test items) to be associated with another vector of information (read diagnostic utilities of item responses). At some determinable point, the expected information gain can be calculated precisely if the elements of both vectors are known (De Groot, 1970, p. 356 ff). In the present instance, the elements in the latter vector can be represented probabilistically.

Each diagnostic hypothesis carries a certain a priori probability of being true, which varies from one hypothesis to another. An efficient diagnostic testing process accumulates evidence to help differentiate between hypotheses by reevaluating the belief probabilities of each of the diagnostic hypotheses after each

response. At some discrete point in the process the probability of one of the working hypotheses will become sufficiently high to justify a report of it as a probable diagnosis. Bayes theorem is used to aggregate the evidence.

The following notation allows demonstration of diagnostic assessment of mastery within a specified curricular subdomain.

(a) Diagnostic hypotheses

H_0 : The student has mastered the subdomain

H_1 :
H_2 :
H_3 : } The student has not mastered the subdomain due to one of five specified learning gaps, misconceptions or misunderstandings.
H_4 :
H_5 :

(b) Probabilities

P_i : the probability that Hypothesis H_i is correct for a particular student, given that one and only one of $[H_0 \ldots H_5]$ is correct.

$P_i^{(k)}$ is the probability after k items have been attempted

(c) Distractors and flags

Suppose that each item has four choices and the "event" that the subject chooses the first one on item k is coded as:

$\underline{X}_k = [1,0,0,0]$; $x_{jk} = 1$ if alternative j on item k is selected, and = 0 otherwise.

Each alternative is flagged to one and only one of the hypotheses; $d_{jk} = i$ if the jth alternative on item k is flagged to hypothesis i.

The prior probabilities specify that in the absence of any specific evidence, each diagnostic hypothesis and the mastery condition are equally likely. Until empirical studies establish the mathematical basis for a more sophisticated Bayesian model, the following will be used:

$$P_i^{(o)} = 1/i_{n=0.167} \quad \text{for } i = 0 \ldots 5$$

The conditional probabilities indicate the strength of the diagnostic information provided by a single test item. Although empirical studies will be necessary to establish this characteristic for any particular type of item, past experience suggests that the probability of selecting the response flagged to a particular and true hypothesis will be somewhere in the range 0.4 and 0.8. A starting value of 0.55 would thus seem to be fairly conservative.

$\text{Prob.} [\underline{X}_k \mid H_i]$ = 0.55 if the chosen alternative is flagged to hypothesis i (i.e., if $x_{jk} = 1$ and $d_{jk} = i$ for some of j).

= 0.15 if one of the rejected alternatives is flagged to hypothesis i (i.e., if $x_{jk} = 0$ and $d_{jk} = 1$ for some j).

= 0.25 if, for this item, none of the distractors are flagged for hypothesis i (i.e., if $d_{jk} = i$ for all of j).

Bayes theorem, now expressed as

$$P_i^{(k)} = \frac{P_i^{(k-1)} \cdot \text{Prob.}[X \mid H_i]}{P_j^{(k-1)} \cdot \text{Prob.}[X \mid H_j]}$$

allows an algorithm to be established for the sequencing of test materials and through such alogrithm the basis for forming a decision as to whether to continue testing. Within an subdomain:

for the first item: Select at random from the full set.

for the second item: (a) Identify all hypotheses not covered in the preceding item.

(b) Select at random from items that include all hypotheses so identified.

for the third item: (a) Identify all hypotheses not covered twice in the preceding items.

(b) Select an item which covers as many of these as possible.

for the fourth item: (a) Identify all hypotheses not covered at least twice in the first three items.

(b) Identify the hypothesis with the greatest P-value.

(c) Select an item to cover hypotheses identified in (a) and (b) above.

Discontinue testing when P_i reaches a confidence level of at least 0.8 for some i. Hypothesis H_i is then reported.

These rules are concise and straightforward. If the student responds consistently, they should lead rapidly to the identification of the appropriate diagnosis.

b.) Illustration of probabilistic differentiation

Consider a sequence of six multiple choice items whose responses

have been flagged to indicate the relevant hypothesis according to the following table. (H_0 denotes the correct answer in each case.)

	ITEM					
Response	1	2	3	4	5	6
A	H_0	H_1	H_4	H_1	H_0	H_2
B	H_1	H_0	H_5	H_2	H_3	H_0
C	H_2	H_3	H_0	H_3	H_2	H_4
D	H_3	H_5	H_1	H_0	H_1	H_5

The prior probabilities of each hypothesis are set equal to 0.167. Response A to item 1 provides some evidence of mastery of the subdomain by the examinee. Bayes theorem combines this evidence with the prior probabilities to give a new probability of mastery:

$$p_0^{(1)} = \frac{0.167 \times 0.55}{0.167 \times 0.55 + 0.167 \times 0.15 + \ldots} = 0.367$$

The probabilities of the other hypotheses are similarly recalculated, and recorded in a table which gives the probabilities of the various hypotheses and the responses selected on successive items. If the subject's response to item 2 was B and to item 3 was C, the following table results.

	Calculated probabilities					
	P_0	P_1	P_2	P_3	P_4	P_5
Prior Values	.167	.167	.167	.167	.167	.167
Item 1 - Response A	.367	.100	.100	.100	.167	.167
Item 2 - Response B	.624	.046	.076	.046	.128	.077
Item 3 - Response C	.834*	.016	.046	.027	.046	.028

After three items testing could be discontinued since the probability of hypothesis H_0 (= mastery) has risen over 0.8.

Next consider a subject who gives a somewhat less consistent pattern of responses. He chooses the alternative flagged for H_1 except on item 2 where his response is flagged for H_5.

Calculated probabilities

	P_0	P_1	P_2	P_3	P_4	P_5
Prior values	.167	.167	.167	.167	.167	.167
Item 1 - Response B	.100	.367	.100	.100	.167	.167
Item 2 - Response D	.061	.226	.102	.061	.171	.377
Item 3 - Response D	.035	.484	.010	.059	.100	.220
Item 4 - Response A	.013	.709	.039	.023	.066	.146
Item 5 - Response D	.003	.858*	.012	.007	.036	.080

In this case we discontinue testing after five items and report H_1. Finally, consider a subject who chooses the response appropriate to H_5 when one is available, but guesses when one is not.

Calculated probabilities

	P_0	P_1	P_2	P_3	P_4	P_5
Prior values	.167	.167	.167	.167	.167	.167
Item 1 - Response B	.090	.367	.100	.100	.167	.167
Item 2 - Response D	.061	.226	.102	.061	.171	.377
Item 3 - Response B	.028	.107	.080	.048	.080	.654
Item 4 - Response D	.065	.068	.051	.030	.085	.698
Item 5 - Response B	.040	.042	.031	.068	.088	.727
Item 6 - Response D	.013	.023	.010	.037	.029	.886*

In this case, all six items are needed before a hypothesis reaches the specified confidence level.

Note that in each of three cases above, the subject responded with a fair degree of consistency. Subjects who respond

inconsistently will need to be given more items before any hypothesis is established. In practice, if this can not be done by ten items then it may be best to report this fact and move on to another area.

Note also that in the above examples, the six items were attempted in a fixed order, in general not the most efficient procedure. For example, after the third subject had attempted three items, H_5 was clearly established as the most probable hypothesis. It would have been better to then administer item 6 (which relates to H_5) rather than items 4 and 5 (which do not). If this had been done (and response D was still selected) testing could have been terminated immediately.

REFERENCES

Box, G.E.P., & Tiao, G.C. Bayesian inference in statistical analysis. Reading MA, Addison-Wesley, 1973.

DeGroot, M.H. Optimal statistical decisions. New York, McGraw-Hill, 1970.

IEA, Tests on understanding the nature of science. Stockholon: International Association for the Evaluation of Education Achievement, 1969.

Novick, M.R. Bayesian considerations in educational information systems. Iowa City, Iowa: American College Testing Program, ACT Research Report #38, 1970.

Novick, M.R., Jackson, P.H., Thayer, D.T., & Cole, D.S. Application of Bayesian methods to the prediction of educational performance. Iowa: 1971 American College Testing Program Research Report #42.

Tamir, P. & Kemp, R.F. Cognitive preference styles across three science disciplines. Science Education, 1978, 62, 143-152.

Tatsuoka, K.K., Birenbaum, M., Tatsuoka, M.M., & Baillie, R. A psychometric approach to error analysis on response patterns. (Research Report 80-3). Urbana, Ill.: University of Illinois, Computer-based Education Research Laboratory, February 1980.